Numer. Math. 75: 421–445 (1997)

Numerische
Mathematik
© Springer-Verlag 1997
Electronic Edition

Accuracy and conservation properties in numerical integration: the case of the Korteweg-de Vries equation

J. de Frutos, J.M. Sanz-Serna

Departamento de Matemática Aplicada y Computación, Universidad de Valladolid, Valladolid, Spain; e-mail: frutos@cpd.uva.es and sanzserna@cpd.uva.es

Received November 21, 1994 / Revised version received July 17, 1995

Summary. When numerically integrating time-dependent differential equations, it is often recommended to employ methods that preserve some of the invariant quantities (mass, energy, etc.) of the problem being considered. This recommendation is usually justified on the grounds that conservation of invariant quantities may ensure that the numerical solution possesses some important qualitative features. However there are cases where schemes that preserve invariants are also advantageous in that they possess favourable error propagation mechanisms that render them superior from a quantitative point of view. In the present paper we consider the Korteweg-de Vries equation as a case study. We show rigorously that, for soliton problems and at leading order, the error of conservative schemes consists of a phase error that grows linearly with time plus a complementary term that is bounded in the H^1 norm uniformly in time. For 'general', nonconservative schemes the error involves a linearly growing amplitude error, a quadratically growing phase error and a complementary term that grows linearly in the H^1 norm. Numerical experiments are presented.

Mathematics Subject Classification (1991): 65M12

1. Introduction

The purpose of this paper is to show that, in the numerical integration of evolutionary problems, schemes that preserve invariants of motion may have favourable error propagation mechanisms leading to better accuracy than one may at first have expected. An example will be presented where a conserving second-order scheme gives more accurate results than a nonconserving, third-order scheme, in spite of the fact that the higher order method has smaller local (truncation) errors.

Correspondence to: J.M. Sanz-Serna

The classical analysis [17] of numerical methods for time-dependent, ordinary or partial differential equations is based on the ideas of stability, consistency and convergence. Roughly speaking, consistency means small local errors and stability means that local errors do not build up catastrophically. Together, consistency and stability yield convergence: small (global) errors. However it is clear that there are useful theoretical properties of a method beyond its consistency, stability and convergence. Here we are interested in *conserved quantities* (first integrals): the differential equations being integrated may possess one or several quantities (mass, energy, etc.) that are conserved in the true evolution and it is reasonable to demand that the numerical scheme also preserves those quantities. Several reasons are usually invoked for using schemes with such conservation properties. In a recent paper [9], C.W. Gear writes “In some cases, failure to maintain certain invariants leads to physically impossible solutions”. In other cases conservation quantities are deemed important to avoid spurious blow-up of the numerical solution. In a classical paper [1], Arakawa writes “If we can find a finite difference scheme which has constraints analogous to the integral constraints of the differential form, the solution will not show the false ‘noodling’, followed by computational instability”.

Since a quantitavely accurate numerical solution cannot show ‘spurious blow-up’ or ‘false noodling’, it follows that the foregoing remarks are meant to apply to integrations in intervals $0 < t < t_{max}$ so long, relatively to the step-size Δt being used, that the numerical solution deviates significantly from the theoretical solution. Thus, it is often believed that, as t becomes large for given Δt, conservative methods go quantitatively wrong but may stay qualitatively acceptable, while nonconservative numerical solutions turn useless from both the quantitative and the qualitative viewpoints.

However such an assessment of the merits of conservative schemes is too severe. In actual fact, in many cases, conservative schemes have better error propagation mechanisms that render them superior from a quantitative point of view. In such cases, conservative algorithms should be preferred even for computations where the numerical solution remains close to the theoretical solution. An instance is presented in [3]. It is shown there that, when integrating the two-body problem with some conservative schemes (including symplectic algorithms that automatically conserve a modified energy), the leading term of the global error grows linearly with t, while for ‘general’ schemes the growth is quadratic. This makes conservative methods more efficient than general methods when accurate solutions are needed. A fuller treatment of these ideas in the case of periodic solutions of *ordinary differential equations* may be seen in [4]. It should be pointed out that the techniques in [3] or [4] are very different from those used here.

In the present paper we use the Korteweg-de Vries (KdV) equation

$$u_t + uu_x + u_{xxx} = 0, \quad -\infty < x < \infty, \quad t > 0, \tag{1}$$

as a model case, but our analysis may be extended to other, not necessarily integrable, equations (see the final section). After presenting the numerical methods considered (Sect. 2) and the KdV equation (Sect. 3), we analyse theoretically the behaviour of the numerical solutions for the case of soliton solutions (Sect. 4). It turns out that schemes that preserve the integrals of the solution and the solution squared behave much better than ‘general’ schemes. We show rigorously that, for soliton problems and at leading order, the error of conservative schemes consists of a phase error that grows linearly with time plus a complementary term that is bounded in the H^1 norm uniformly in time. For ‘general’, nonconservative schemes the error involves a linearly growing amplitude error, a quadratically growing phase error and a complementary term that grows linearly in the H^1 norm. These analytical findings are proved only for single soliton solutions but are nevertheless relevant because all other solutions asymptotically give rise to solitons. Numerical illustrations involving single soliton solutions and interactions of solitons are presented in Sect. 5. The advantages of conservation are clearly borne out, particularly so in the case of long-time integrations. The final Sect. 6 contains some concluding remarks.

The main observation in the paper is that, if we look at the local error of a numerical method as a vector in a suitable phase space, then conservation properties imply constraints for the direction of the local error. When local errors build up to give rise to the global error, their directions are not irrelevant: there are harmful directions that lead to faster error accumulation. In many instances, the local error of a conservative scheme has a *direction* that renders it relatively harmless and this gives the scheme an advantage. These features are not captured by standard convergence analyses, which just take into account the *size* of the local error.

A preliminary study [7] of the issues addressed here was presented at the 1993 Dundee conference. The unpublished report [8] contains some proofs not included in [7]. The material in [7] and [8] is based on nonrigorous soliton perturbation results [13], [14] and on Benjamin’s classical soliton stability theorem [2]. In the present paper we use recent stability results due to Pego and Weinstein [16]; these results are more powerful than those in [2], [13], [14] and correspondingly our conclusions here are stronger than the conclusions of our earlier work [7], [8].

In this paper ‘conservative scheme’ refers to a scheme that preserves an invariant quantity. There are other ideas of conservation in numerical analysis. The concept of *symplectic algorithm* for Hamiltonian problems [18] relates to conservation, not of a quantity, but of a *differential form*. In numerical hyperbolic problems retaining in the scheme the *conservation format* of the differential equation is required to obtain the correct weak solution [15].

2. Numerical methods

2.1. Preliminaries

We consider semidiscrete (discrete t, continuous x) numerical methods for (1). It is best to present the algorithms as applied to a general evolution equation

$$u_t = f(u), \tag{2}$$

where, for each value of the time t, the state $u(t)$ is an element of a real vector space V. For ordinary differential equations, V is finite dimensional; for partial differential equations, V is an infinite dimensional space consisting of functions of the space variables. The notation ϕ_t refers to the the time t flow of (2), so that $\phi_t(\alpha)$ is the value at time t of the solution of (2) with initial value α at time 0. For simplicity, our treatment in the remainder of this section is only *formal;* in particular we shall not spell out the hypotheses required for ϕ_t to be well defined, the choice of norm in V, etc. The lack of rigour in this section does not affect the remainder of the paper: propositions proved here *motivate* later developments, but are not actually *used* in the derivation of the main results.

The system (2) is integrated by a one-step method

$$U^{n+1} = \Phi_{\Delta t}(U^n), \tag{3}$$

where Δt denotes the time step and U^n is the numerical solution at time level $t_n = n\Delta t$. Obviously the mapping $\Phi_{\Delta t}$ should approximate the true evolution given by $\phi_{\Delta t}$. The local error (at a state $u_0 \in V$) is, by definition,

$$L_{\Delta t}(u_0) = \Phi_{\Delta t}(u_0) - \phi_{\Delta t}(u_0).$$

If p denotes the order of the method, then $L_{\Delta t}(u_0) = O(\Delta t^{p+1})$. For most methods used in practice, the local error possesses an asymptotic expansion

$$L_{\Delta t}(u_0) = \Delta t^{p+1}\ell_{p+1}(u_0) + \Delta t^{p+2}\ell_{p+2}(u_0) + \dots . \tag{4}$$

Note that $L_{\Delta t}$, and therefore ℓ_{p+1}, ℓ_{p+2}, ... are mappings defined in V with values in V. The mapping $L_{\Delta t}$ depends on the parameter Δt, but the ℓ_{p+k} do not.

Now assume that the system (2) with initial condition $u(0) = \alpha$ is integrated by the method (3). If $u(\cdot)$ denotes the true solution and $\{U_n\}$ the corresponding numerical solution ($U_0 = \alpha$), the global error $U_n - u(t_n)$ possesses a formal asymptotic expansion

$$U_n - u(t_n) = \Delta t^p e_p(t_n) + \Delta t^{p+1} e_{p+1}(t_n) + \dots . \tag{5}$$

Here e_p, e_{p+1}, ... are V-valued functions of t, independent of Δt. These functions are found (see [11] Chapter II.8) by solving linear initial value problems (variational equations)

$$\frac{d}{dt}e_{p+k} = f'(u(t)) \cdot e_{p+k} + s_{p+k}(t), \quad e_{p+k}(0) = 0, \quad k = 0, 1, 2, \dots . \tag{6}$$

The symbol $f'(u(t))$ refers to the derivative of f evaluated at the state $u(t)$; a linear operator in V. The source terms $s_{p+k}(t)$ are functions of t with values in V; they are computable in terms of the coefficients ℓ_{p+k} in (4). The source corresponding to the leading order is the leading term of the local error evaluated at the true solution, i.e.,

$$s_p(t) = \ell_{p+1}(u(t)). \tag{7}$$

The expression for subsequent sources is more involved, for instance, assuming that $p \geq 2$,

$$s_{p+1}(t) = \ell_{p+2}(u(t)) - \frac{1}{2} f'(u(t)) \cdot \ell_{p+1}(u(t)) - \frac{1}{2}\frac{d}{dt}\ell_{p+1}(u(t)),$$

see [11] Chapter II.8, Exercise 1 (there is a misprint in this exercise in the first edition of [11]).

2.2. Conserved quantities

We now assume that I is a real-valued function defined in V that is conserved by solutions of (2), i.e., $I(\phi_t(\alpha)) = I(\alpha)$ for all real t and all $\alpha \in V$. By differentiation with respect to t, it follows easily that

$$\forall u_0 \in V, \quad I'(u_0) \cdot f(u_0) = 0. \tag{8}$$

Conversely, if (8) holds for a function I, then I remains constant along solutions of (2). Note that (8) can be rephrased by saying that, at each state u_0, the vector $f(u_0)$ must lie in the kernel of the linear functional $I'(u_0)$. If V is an inner-product space, with inner product $< \cdot, \cdot >$, then $I'(u_0) \cdot v \equiv < g(u_0), v >$, for a suitable vector $g(u_0)$ (the gradient vector of I at u_0); in this case (8) demands that $f(u_0)$ should be orthogonal to $g(u_0)$.

A step from the state $u_0 \in V$ with the method (3) changes I by an amount (see (4))

$$\begin{aligned} I(\Phi_{\Delta t}(u_0)) - I(u_0) &= I(\Phi_{\Delta t}(u_0)) - I(\phi_{\Delta t}(u_0)) \\ &= \Delta t^{p+1} I'(u_0) \cdot \ell_{p+1}(u_0) + O(\Delta t^{p+2}), \end{aligned} \tag{9}$$

which in general is $O(\Delta t^{p+1})$. However for some methods the change can be $O(\Delta t^{p+k})$ for $k > 1$ and, of course, there are also instances where for a given equation, a given conserved quantity and a given numerical method, exact conservation holds, i.e., $I(\Phi_{\Delta t}(u_0)) = I(u_0)$ for all Δt and all $u_0 \in V$. For instance, all practical methods exactly conserve *linear* invariants, see e.g. [19], and a class of Runge-Kutta methods, including the midpoint rule, exactly conserve *quadratic* invariants [5]. When I is not linear or quadratic, *ad hoc* special schemes have been constructed to achieve exact conservation; the literature on this topic is quite extensive and cannot be reviewed here, cf. [18].

From (9) we can state:

Lemma 1. *A step from a state $u_0 \in V$ with a method of order p changes the conserved quantity I by an $O(\Delta t^{p+1})$ amount. The change is $O(\Delta t^{p+2})$ for all states u_0 and all time steps Δt if and only if*

$$\forall u_0 \in V, \quad I'(u_0) \cdot \ell_{p+1}(u_0) = 0. \tag{10}$$

In particular (10) holds for methods that exactly conserve I.

Note that (10) demands that $\ell_{p+1}(u_0)$ be in the kernel of $I'(u_0)$, or in the inner product case that $\ell_{p+1}(u_0)$ be orthogonal to the gradient $g(u_0)$ of I at u_0. Thus conservation properties are linked to the *direction* in the phase space V of the local error.

Conditions similar to (10) guaranteeing $O(\Delta t^{p+k})$, $k = 3, 4, \ldots$, changes may be obtained, but they are cumbersome. Therefore we leave the investigation of the relations between the local error expansion and conservation properties and take up the study of the connections between conservation properties and *global* errors. We look at the quantity $I(U^n) - I(U^0) = I(U^n) - I(u(t_n))$, where $\{U_n\}$ and $u(\cdot)$ are the numerical and true solutions corresponding to the initial state α. This quantity is both the error in the numerical computation of $I(u(t_n))$ and the spurious growth in I due to numerical integration from $t = 0$ to $t = t_n$. From (5):

$$\begin{aligned} I(U^n) - I(U^0) &= I(U^n) - I(u(t_n)) \\ &= I'(u(t_n)) \cdot [U^n - u(t_n)] + O(\Delta t^{2p}) \\ &= \Delta t^p I'(u(t_n)) \cdot e_p(t_n) + \cdots + \Delta t^{2p-1} I'(u(t_n)) \cdot e_{2p-1}(t_n) + O(\Delta t^{2p}). \end{aligned}$$

Hence, as $\Delta t \to 0$ with t_n fixed, the error $I(U^n) - I(u(t_n))$ is, in general, $O(\Delta t^p)$. To obtain an $O(\Delta t^{p+1})$ behaviour, one would need $I'(u(t)) \cdot e_p(t) = 0$ for all t, etc. In this connection we have (cf. [3], Lemma 2):

Lemma 2. *Let k be an integer $1 \le k \le p$. Assume that, for the initial condition α, the p-th order method (3) has errors $I(U^n) - I(u(t_n))$ that are $O(\Delta t^{p+k})$ for all t_n. Then the source terms in the variational equation (6) satisfy*

$$\forall t, \quad I'(u(t)) \cdot s_{p+j}(t) \equiv 0, \quad j = 0, \ldots, k-1 \tag{11}$$

These relations are valid in particular for methods that conserve I exactly.

Proof. Differentiate $I'(u(t)) \cdot e_{p+j}(t) = 0$ with respect to t and use (2) and (6) to obtain

$$I''(u(t)) \cdot [f(u(t)), e_{p+j}(t)] + I'(u(t)) \cdot f'(u(t)) \cdot e_{p+j}(t) + I'(u(t)) \cdot s_{p+j}(t) = 0,$$

where $I''(u(t)) \cdot [\cdot, \cdot]$ is the second derivative of I evaluated at $u(t)$, a bilinear symmetric operator. Now differentiation in (8) with respect to u_0 leads to

$$\forall u_0 \in V, \quad I''(u_0) \cdot [\cdot, f(u_0)] + I'(u_0) \cdot f'(u_0) = 0$$

and (11) follows readily. □

The case $j = 0$ in (11), reduces, in view of (7), to $I'(u(t)) \cdot \ell_{p+1}(u(t)) = 0$. Hence if the global errors in I are $O(\Delta t^{p+1})$ for all initial conditions $\alpha \in V$, then (10) holds as expected.

3. The Korteweg-de Vries equation

3.1. The nonlinear equation

Among the many remarkable properties of (1) we focus on two: the existence of conservation laws and the existence of solitons.

There is an infinite number of conservation laws for the initial value problem for (1), but we only need the first two. For smooth solutions the following quantities, that we shall respectively call *mass* and *energy*, do not vary with t:

$$I_1(u) = \int_{-\infty}^{\infty} u(x,t)\,dx, \tag{12}$$

$$I_2(u) = \int_{-\infty}^{\infty} u^2(x,t)\,dx. \tag{13}$$

(In many physical applications the integral I_2 is referred to as momentum. We here use the word energy that is standard in mathematics.)

Introduce the following real function of w of two variables ξ (real) and A (positive)

$$w(\xi, A) = A\,\mathrm{sech}^2 \frac{\sqrt{3A}}{6}\xi; \tag{14}$$

the KdV equation possesses, for each choice of real constants $A > 0$ and μ, the travelling wave solution

$$u(x,t) = w(x - ct - \mu, A), \tag{15}$$

where μ determines the location of the wave at $t = 0$ and the speed of propagation $c = A/3$ is a function of the wave amplitude A. It is important to note that the taller the wave the faster it travels.

For the soliton (15) the values of the conserved quantities (12)–(13) are

$$I_1 = 4\sqrt{3}A^{1/2}, \qquad I_2 = \frac{8\sqrt{3}}{3}A^{3/2}. \tag{16}$$

3.2. Linearization

We now fix values A_0 and μ_0 for the soliton parameters and study the soliton $w(x - c_0t - \mu_0, A_0)$, $c_0 = A_0/3$. To simplify the notation, we introduce the real function w_0 of a real variable such that

$$w_0(x - c_0t) = w(x - c_0t - \mu_0, A_0).$$

Thus the initial condition $u(x,0) = \alpha(x)$ given by

$$\alpha(x) = w_0(x), \qquad -\infty < x < \infty, \tag{17}$$

gives rise to the KdV solution $w_0(x - c_0t)$. For an initial condition $\tilde{\alpha}(x) = \alpha(x) + \epsilon\delta(x)$ close to α, the solution is, formally, $\tilde{u}(x,t) = w_0(x - c_0t) + \epsilon e(x,t) +$

$O(\epsilon^2)$. Substituting $\tilde{u}$ in the KdV equation and equating powers of ϵ, it is easily concluded that e satisfies

$$(18)\quad e_t + w_0(x - c_0 t)e_x + w_0'(x - c_0 t)e + e_{xxx} = 0, \quad -\infty < x < \infty, \quad t > 0,$$

a differential equation that has to be supplemented with the initial condition

$$e(x,0) = \delta(x), \qquad -\infty < x < \infty.$$

The equation (18) provides the *linearization* of the KdV equation around the soliton solution w_0. In the abstract notation of the preceding section, we would write the KdV equation as $u_t = f(u)$ and the soliton being studied as $u(t) = w_0(\cdot - c_0 t)$; then (18) is nothing but the homogeneous variational equation $de/dt = f'(u(t)) \cdot e$. For this reason, the variational equations (6) that one has to solve to find the coefficients of the global error expansion of a numerical method are of the form

$$(19)\quad e_t + w_0(x - c_0 t)e_x + w_0'(x - c_0 t)e + e_{xxx} = s(x,t), \quad -\infty < x < \infty, \quad t > 0,$$

where s is the corresponding source.

Let us now investigate the linear differential equations (18) and (19). We begin by presenting two particular (classical) solutions of (18):

$$(20)\qquad w_0'(x - c_0 t), \qquad z_0(x - c_0 t) - t w_0'(x - c_0 t),$$

where z_0 is the function defined by

$$z_0(\xi) = 3\frac{\partial w}{\partial A}(\xi - \mu_0, A_0).$$

(A referee has pointed out that these solutions are sometimes called *zero modes* or *Goldstone modes* in the physics literature.) In order to give an interpretation to these solutions, we first observe that the initial condition $\tilde{\alpha}(x) = w_0(x) + \epsilon w_0'(x)$ is of the form $w_0(x + \epsilon) + O(\epsilon^2)$, i.e., *adding* w_0' *to* w_0 *induces a phase shift.* Now the KdV solution corresponding to the initial datum $w_0(x + \epsilon)$ is obviously $w_0(x - c_0 t + \epsilon)$, which in turn is of the form $w_0(x - c_0 t) + \epsilon w_0'(x - c_0 t) + O(\epsilon^2)$. Therefore $w_0'(x - c_0 t)$ has to satisfy the variational equation (18). On the other hand, consider the initial condition $\tilde{\alpha}(x) = w_0(x) + \epsilon z_0(x)$. This satisfies, by definition of z_0, $\tilde{\alpha}(x) = w(x - \mu_0, A_0 + 3\epsilon) + O(\epsilon^2)$, i.e., *adding* z_0 *to* w_0 *induces a change in amplitude.* The KdV solution with initial datum $w(x - \mu_0, A_0 + 3\epsilon)$ is obviously $w(x - (c_0 + \epsilon)t - \mu_0, A_0 + 3\epsilon)$, which is of the form $w_0(x - c_0 t) - \epsilon t w_0'(x - c_0 t) + \epsilon z_0(x - c_0 t) + O(\epsilon^2)$. This accounts for the second solution in (20).

The preceding discussion may have been presented without any algebra. A shift of length ϵ of a soliton profile at time $t = 0$ has at all later times the effect of shifting the soliton by the same amount. On the other hand, if the initial perturbation is an increase of 3ϵ units in the amplitude A, the result is a new, taller soliton whose speed is ϵ units larger than that of the unperturbed soliton. Hence the perturbed soliton lies ahead of the unperturbed soliton, at a distance ϵt that grows linearly with t.

Let us now turn to the nonhomogeneous variational equation (19). When the source term s is given by $w_0'(x - c_0 t)$ and the initial condition is 0, (19) possesses the solution

$$tw_0'(x - c_0 t). \tag{21}$$

For the source term $z_0(x - c_0 t)$, the solution with trivial initial condition is

$$tz_0(x - c_0 t) - \frac{t^2}{2} w_0'(x - c_0 t). \tag{22}$$

The interpretation of these solutions via Duhamel's principle is not difficult. A steady forcing given by $\epsilon w_0'$ in the KdV equation keeps shifting the soliton w_0 by distance ϵdt in each time interval $[t, t + dt]$. These infinitesimal shifts combine to yield a shift of ϵt units in the evolution from 0 to t. On the other hand, forcing steadily with the source ϵz_0 in the evolution from 0 to t results in an $O(\epsilon t)$ increase in amplitude. Accordingly, the soliton velocity also increases by an $O(\epsilon t)$ amount, which in a time interval of length t induces a change in location of order $O(\epsilon t^2)$.

3.3. The homogeneous linearized equation

For a deeper study of (18) it is convenient to use the moving coordinates $X = x - c_0 t$, $T = t$. In the moving frame of reference, the soliton $w_0(x - c_0 t)$ becomes an equilibrium (i.e., T-independent) solution $w(X)$ of the KdV equation.

As a function $E(X, T) = E(x - c_0 t, t) = e(x, t)$ of the new variables, a solution e of (18) satisfies

$$E_T - c_0 E_X + w_0(X) E_X + w_0'(X) E + E_{XXX} = 0, \qquad -\infty < X < \infty, \quad T > 0;$$

we rewrite this differential equation in the form

$$E_T = \partial_X \mathscr{L} E, \tag{23}$$

where $\mathscr{L}$ is the second-order linear operator

$$\mathscr{L} E = -E_{XX} + (c_0 - w_0(X))E.$$

Pego and Weinstein [16] study (23) in a weighted Sobolev space H_a^1. This consists of all the functions $v(X)$ such that $\exp(aX)v(X)$ lies in the standard space H^1. Furthermore

$$\|v\|_{H_a^1} = \|\exp(aX)v\|_{H^1}.$$

Here a is fixed in the range $0 < a < \sqrt{A_0}/3$. Note that the function $w_0(X)$ and all its derivatives belong to H_a^1, because, according to (14), $w_0(X)$ behaves as $\exp(-\sqrt{A_0/3}|X|)$ as $|X| \to \infty$.

The operator $\partial_X \mathscr{L}$ generates a strongly continuous semigroup of operators $\exp(T\partial_X \mathscr{L})$ in H_a^1, so that the solution of (23) with initial condition $\alpha \in H_a^1$ is $E = \exp(T\partial_X \mathscr{L})\alpha$. Furthermore:

Lemma 3. *(Pego and Weinstein [16] Proposition 2.8) In the space* H^1_a, 0 *is the unique eigenvalue of the operator* $\partial_X \mathscr{L}$. *The geometric multiplicity of this eigenvalue is 1 and, moreover,* $\ker(\partial_X \mathscr{L}) = \mathrm{span}(w_0')$. *The algebraic multiplicty is 2, i.e., the generalized kernel* $\cup_{k=1}^{\infty} \ker((\partial_X \mathscr{L})^k)$ *of* $\partial_X \mathscr{L}$ *has dimension 2. This generalized kernel is spanned by the functions* $w_0' \in \ker(\partial_X \mathscr{L})$ *and* $z_0 \in \ker((\partial_X \mathscr{L})^2)$. *More precisely* $\partial_X \mathscr{L} z_0 = -w_0' \in \ker(\partial_X \mathscr{L})$.

Therefore, in the basis w_0', z_0, the restriction of $\partial_x \mathscr{L}$ to its generalized kernel is expressed by a Jordan matrix

$$\begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix}$$

and (23) has solutions $\beta(T) w_0'(X) + \gamma(T) z_0(X)$ where $d\beta/dT = -\gamma$ and $d\gamma/dT = 0$, i.e., solutions of the form $(-c_2 T + c_1) w_0'(X) + c_2 z_0(X)$. In terms of the original variables x and t we recover the linear combinations of the particular solutions (20).

There is a natural projection P of the space H^1_a onto the generalized kernel:

$$P v =< v, \eta_1 > w_0' + < v, \eta_2 > z_0,$$

where $< \cdot, \cdot >$ denotes the standard inner product

$$< v_1, v_2 >= \int_{-\infty}^{\infty} v_1(X) v_2(X)\, dX$$

and η_1, η_2 are a basis of the generalized kernel of the adjoint operator of $\partial_X \mathscr{L}$ chosen in such a way that

$$\begin{aligned} < w_0', \eta_1 >= 1, \quad & < w_0', \eta_2 >= 0, \\ < z_0, \eta_1 >= 0, \quad & < z_0, \eta_2 >= 1 \end{aligned}$$

(biorthogonality). Explicitly:

$$\begin{aligned} \eta_1(X) &= -\frac{1}{2A_0} \left[\tanh Z + \tanh^2 Z + Z \,\mathrm{sech}^2 Z \right], \\ \eta_2(X) &= \frac{\sqrt{3A_0}}{18} \,\mathrm{sech}^2 Z, \end{aligned}$$

where

$$Z = \frac{\sqrt{3A_0}}{6} X.$$

If $Q = I - P$ denotes the complementary projection of P, then $Q(H^1_a)$ is a complementary subspace of the generalized kernel $P(H^1_a)$ of $\partial_X \mathscr{L}$. It is well known from operator theory that $Q(H^1_a)$ is invariant by $\partial_X \mathscr{L}$, in the sense that, if $v \in Q(H^1_a)$ is in the domain of $\partial_X \mathscr{L}$, then $\partial_X \mathscr{L} v \in Q(H^1_a)$. As a consequence $\exp(T \partial_X \mathscr{L})\alpha$ remains in $Q(H^1_a)$ for all T if $\alpha \in Q(H^1_a)$. As noticed above, solutions in $P(H^1_a)$ are related to perturbations of the soliton amplitude and phase. Solutions in $Q(H_a^1)$ represent perturbations that take the soliton being studied out of the two-dimensional family of soliton solutions.

The restriction of $\partial_X \mathscr{L}$ to $Q(H_a^1)$ has a spectrum consisting of a curve that lies in the left half plane and is bounded away from the imaginary axis ([16], Proposition 2.5). This suggests a decaying behaviour for the solutions and, in fact, Pego and Weinstein show the following result:

Lemma 4. *There are positive constants b and C such that for all initial conditions $\alpha \in Q(H_a^1)$ and all positive times T*

$$\| \exp(T\partial_X \mathscr{L})\alpha\|_{H_a^1} \leq C \exp(-bT)\|\alpha\|_{H_a^1}. \tag{24}$$

For the purposes of this paper the weighted space H_a^1 is only an auxiliary technical tool; the main results in Sect. 4 correspond to the standard H^1 norm. It is easy to check that $\partial_X \mathscr{L}$ generates a strongly continuous semigroup in H^1. Moreover we have the following estimate:

Lemma 5. *There exists a positive constant C such that for all initial conditions $\alpha \in Q(H_a^1) \cap H^1$ and all positive times T*

$$\| \exp(T\partial_X \mathscr{L})\alpha\|_{H^1} \leq C(\|\alpha\|_{H^1} + \exp(-bT)\|\alpha\|_{H_a^1}). \tag{25}$$

Proof. Solutions E in H^1 of (23) conserve the Hamiltonian functional ([18], Sect. 14.7)

$$J(E) = \int_{-\infty}^{\infty} \left[\frac{1}{2}c_0 E(X,T)^2 - \frac{1}{2}w_0'(X)E(X,T)^2 + \frac{1}{2}E_X(X,T)^2\right] dX.$$

We can write (the value of C is not the same at each occurrence)

$$\begin{aligned}
\|E(\cdot,T)\|_{H^1}^2 &\leq CJ(E(\cdot,T)) + \frac{1}{2}\left|\int_{-\infty}^{\infty} w_0'(X)E(X,T)^2\, dX\right| \\
&= CJ(E(\cdot,0)) \\
&\quad + \frac{1}{2}\left|\int_{-\infty}^{\infty} (w_0'(X)\mathrm{e}^{-aX})(\mathrm{e}^{aX}E(X,T))E(X,T)\, dX\right| \\
&\leq CJ(E(\cdot,0)) + C\|E(\cdot,T)\|_{H_a^1}\|E(\cdot,T)\|_{H^1} \\
&\leq CJ(E(\cdot,0)) + \frac{1}{2}\|E(\cdot,T)\|_{H^1}^2 + C\|E(\cdot,T)\|_{H_a^1}^2.
\end{aligned}$$

Here we have taken into account that $w_0'(X)\mathrm{e}^{-aX}$ is bounded. It is now sufficient to estimate $J(E(\cdot,0))$ in terms of $\|E(\cdot,0)\|_{H^1}$ and to use (24). □

3.4. The nonhomogeneous linearized equation

We only study the equation (19) in the particular case where the source $s(x,t)$ is of the form $s(x-c_0t)$, so that in the moving coordinates, the equation reads

$$E_T = \partial_X \mathscr{L} E + s(X). \tag{26}$$

The following result is easily checked by substitution of (27) in (26).

Lemma 6. *If $s \in H^1_a$, then the solution E of (26) with initial condition $E(T=0) = 0$ is*

$$\begin{aligned} E(X,T) \quad = \quad & < s,\eta_1 > Tw_0' + < s,\eta_2 > \left(Tz_0 - \frac{T^2}{2} w_0' \right) \\ & + \int_0^T \mathrm{e}^{(T-\tau)\partial_X \mathscr{L}} Qs \, d\tau, \end{aligned} \tag{27}$$

Note that the first two terms in the right hand side of (27) are a linear combination of the particular solutions (21)–(22) we discussed above. These terms provide the projection of E onto the genera lized kernel of the operator $\partial_X \mathscr{L}$ and therefore represent changes in soliton phase and soliton amplitude.

In some cases, the integral in (27), i.e, the projection of E onto $Q(H^1_a)$, can be computed explicitly:

Lemma 7. *In the situation of the preceding lemma, assume that Qs lies in the range of the operator $\partial_X \mathscr{L}$ in H^1_a, so that $Qs = \partial_X \mathscr{L} \sigma$ for some $\sigma \in H^1_a$. Then*

$$\int_0^T \mathrm{e}^{(T-\tau)\partial_X \mathscr{L}} Qs \, d\tau = (I - \mathrm{e}^{T\partial_X \mathscr{L}})\sigma. \tag{28}$$

Proof. The expression $-\mathrm{e}^{(T-\tau)\partial_X \mathscr{L}} \sigma$ provides an antiderivative of the integrand. □

The estimate (25) shows that, for $Qs \in H^1$, the integral in (27), having a bounded integrand, grows at most linearly with t in the H^1 norm. The representation (28) implies that, if $\sigma \in H^1$, then the integral actually remains, for all T, *bounded* in the H^1 norm.

4. Main results

Let us now assume that the KdV equation (1) is numerically integrated, with the initial condition (17) we have been considering, by a one-step numerical method of order p. We make the following (reasonable) hypotheses:

(H1) The numerical solution U^n at time $t_n = n\Delta t$ exists, at least for Δt sufficiently small (how small may depend on the value of t_n).

(H2) The numerical solution possesses in H^1 an asymptotic expansion of the form

$$(29) \qquad U^n(x) = w_0\,(x - c_0 t_n) + \Delta t^p e(x, t_n) + \Delta t^p R(x, t_n, \Delta t)\,,$$

where the H^1 function e is independent of Δt and satisfies the variational equation (19) for a suitable source and R is a remainder such that $\|R(\cdot, t, \Delta t)\|_{H^1} \to 0$ as $\Delta t \to 0$.

(H3) The source term mentioned in (H2) is of the form $s(x - c_0 t)$, where s is a $\mathscr{C}^\infty$ function of a real variable. Furthermore, as $|\xi| \to \infty$, $s(\xi)$ behaves as $\mathrm{e}^{-\sqrt{A_0/3}|\xi|}$.

Only the hypothesis (H3) needs some comments. As explained in Sect. 2, at the leading order, the source term in the variational equation is given by $\ell_{p+1}(u(t))$, i.e., by the leading coefficient of the local error evaluated at the theoretical solution. In the present application, the theoretical solution is of class $\mathscr{C}^\infty$, depends on x and t through the combination $x - c_0 t$ and behaves as $\mathrm{e}^{-\sqrt{A_0/3}|x|}$ for $|x|$ large. Therefore (H3) is a most reasonable hypothesis.

We are now in a position to give the main result of the paper.

Theorem 1. *Assume that the hypotheses (H1)–(H3) above hold. Then*

$$\begin{aligned} U^n(x) \quad &= \quad w(x - c_0 t_n - \mu_0 + \Delta t^p \lambda_1 t_n - \frac{\Delta t^p}{2}\lambda_2 t_n^2, A_0 + \frac{\Delta t^p}{3}\lambda_2 t_n) \\ (30) \qquad & \qquad + \Delta t^p \rho(x, t_n) + \Delta t^p \tilde{R}(x, t_n, \Delta t), \end{aligned}$$

where

$$\lambda_1 =< s, \eta_1 >, \qquad \lambda_2 =< s, \eta_2 >,$$

the function ρ is independent of Δt, satisfies $\rho(\cdot, t) \in Q(H^1_a)$ and possesses a bound

$$(31) \qquad \|\rho(\cdot, t)\|_{H^1} \leq Ct,$$

and $\tilde{R}$ is a remainder such that, for each fixed time t, $\|\tilde{R}(\cdot, t, \Delta t)\|_{H^1} \to 0$, as $\Delta t \to 0$.

Proof. By (H3) the source $s(x - c_0 t)$ is such that $s \in H^1_a \cap H^1$. By (27) (rewritten in the (x, t) coordinates), the function $e(x, t)$ in (H2) is of the form

$$\begin{aligned} e(x, t) \quad &= \quad < s, \eta_1 > t w_0'(x - c_0 t) \\ & \qquad + < s, \eta_2 > \left(t z_0(x - c_0 t) - \frac{t^2}{2} w_0'(x - c_0 t) \right) \\ (32) \qquad & \qquad + \rho(x, t), \end{aligned}$$

where, for each t, $\rho(\cdot, t) \in Q(H^1_a)$. The fact that the growth of ρ is at most linear in the H^1 norm was noticed in the remark following Lemma 7. We substitute e from (32) in (29). This gives rise to the combination

$$w_0(x - c_0t) + \Delta t^p < s, \eta_1 > t w_0'(x - c_0t) \\ + \Delta t^p < s, \eta_2 > \left(t z_0(x - c_0t) - \frac{t^2}{2} w_0'(x - c_0t) \right),$$

which, by definition of w_0 and z_0, differs from

$$w(x - c_0t_n - \mu_0 + \Delta t^p \lambda_1 t_n - \frac{\Delta t^p}{2} \lambda_2 t_n^2, A_0 + \frac{\Delta t^p}{3} \lambda_2 t_n)$$

in $O(\Delta t^{2p})$ terms that can be hidden in the remainder. □

From the theorem we see that the numerical solution consists of three components. The first, that we call *modified soliton*, has, at each time t, the *exact* shape of a KdV soliton profile. However the modified soliton has an amplitude $A_0 + (\Delta t^p/3)\lambda_2 t$ that, as the integration proceeds, keeps varying at a steady rate $(\Delta t^p/3)\lambda_2$. Furthermore, the modified soliton has a phase that is an error by an amount $\Delta t^p \lambda_1 t - (\Delta t^p/2)\lambda_2 t^2$ growing *quadratically* with t. (Cf. the discussion following (22).) The amplitude and phase errors in the modified soliton are of the order of Δt^p. The second term in (30), that we call *complementary error*, represents those numerical errors that, while being of the leading order $O(\Delta t^p)$, cannot be interpreted as changes in the soliton amplitude and phase (w lies at each time t in the complementary $Q(H_a^1)$ of the generalized kernel). For instance $\Delta t^p \rho$ may account for the fact that the shape of the numerical solution is not exactly the same as that of a true KdV soliton. Also $\Delta t^p \rho$ may contain a soliton *tail;* an issue that we will discuss when presenting the numerical results. Finally the third term in (30) represents a higher order, $o(\Delta t^p)$ remainder.

When t is large and Δt is so small (relatively to t) that the higher order remainder may be ignored, the dominant contribution to the error is the $O(t^2 \Delta t^p)$ phase error; the amplitude error and the complementary error $\Delta t \rho$ only grow *linearly* with t.

Let us now relate Theorem 1 and the conservation laws. As discussed in Sect. 2, all sensible methods exactly conserve the mass I_1, because this is a linear functional. Accordingly (Lemma 2) we expect that the source $s(x - c_0t)$ be orthogonal, for each t to the gradient of mass. Now the gradient of mass is formally given by the function 1, because

$$I_1(u_0 + \epsilon v) = I_1(u_0) + \epsilon < 1, v > .$$

Therefore, we expect, for any reasonable method, $< 1, s >= 0$, or in other words $\int_{-\infty}^{\infty} s(\xi)\, d\xi = 0$. Similarly the gradient of energy at the soliton $w_0(x - c_0t)$ is the function $2w_0(x - c_0t)$ and hence for methods that conserve exactly energy $< s, w_0 >= 0$. By Lemma 2, the same orthogonality relation is expected to hold for methods of order p that, when integrating the soliton, produce global energy errors of order $O(\Delta t^{p+1})$. For conservative methods the estimates in Theorem 1 can be improved considerably.

Theorem 2. *Assume that the hypotheses (H1)–(H3) above hold and that the source s satisfies the 'conservation properties' $< s, 1 >=< s, w_0 >= 0$. Then*

$$U^n(x) = w(x - c_0 t_n - \mu_0 + \Delta t^p \lambda_1 t_n, A_0) \tag{33}$$
$$+\Delta t^p \rho(x, t_n) + \Delta t^p \tilde{R}(x, t_n, \Delta t),$$

where

$$\lambda_1 =< s, \eta_1 >,$$

the function ρ is independent of Δt, satisfies $\rho(\cdot, t) \in Q(H_a^1)$ and possesses a bound

$$\|\rho(\cdot, t)\|_{H^1} \leq C, \tag{34}$$

and $\tilde{R}$ is a remainder such that, for each fixed time t, $\|\tilde{R}(\cdot, t, \Delta t)\|_{H^1} \to 0$, as $\Delta t \to 0$.

Proof. The functions $w_0(x - c_0 t)$ and η_2 differ in a multiplicative constant, so that the hypothesis $< s, w_0 >= 0$ implies $\lambda_2 = 0$ in (30).

On the other hand, we are going to prove that the projection Qs is in the range of the operator $\partial_X \mathscr{L}$ (in H_a^1), so that Lemma 7 applies, leading to the boundedness of ρ in the H^1 norm.

We begin by showing that

$$< Qs, 1 >= 0.$$

In fact, by definition of $Q = I - P$,

$$< Qs, 1 >=< s, 1 > - < s, \eta_1 >< w_0', 1 > - < s, \eta_2 >< z_0, 1 >;$$

the first and third term in the right hand side vanish because of the hypothesis on s, the second vanishes because $< w_0', 1 >= 0$.

Once we know that the integral of Qs vanishes, it is clear that

$$\Lambda(X) = \int_{-\infty}^{X} (Qs)(\xi)\, d\xi$$

is a smooth function that lies in $H_a^1 \cap H^1$ and such that $\partial_X \Lambda = Qs$. It remains to show that Λ is in the range of $\mathscr{L}$, or, in other words, that Λ is orthogonal to the kernel of the adjoint operator. This kernel is precisely the span of the function w_0'. By integration by parts, $< \Lambda, w_0' >= - < s, w_0 >= 0$ and the proof is ready. □

Note that in this case the modified soliton keeps the correct amplitude A_0 and has a phase error that only grows *linearly* with time. Also the complementary error remains bounded. The most harmful direction that the source term may have is that of the soliton profile w_0 (i.e., that of η_2). It is the component of s in this direction that excites the quadratic growth explained in the discussion following (22). Preservation of energy ensures that s has a vanishing component in this harmful direction.

5. Numerical examples

5.1. Schemes being compared

We consider the family of singly diagonally implicit Runge-Kutta (SDIRK) methods

$$\begin{array}{|cc} \gamma & 0 \\ 1-2\gamma & \gamma \\ \hline \frac{1}{2} & \frac{1}{2} \end{array}\ ,$$

i.e., the methods that integrate (2) according to the recipe

$$U^{n+1} = U^n + \frac{\Delta t}{2}[f(U^*) + f(U^{**})],$$

where the stage vectors U^* and U^{**} are obtained by successively solving the equations

$$\begin{aligned} U^* &= U^n + \Delta t\gamma f(U^*), \\ U^{**} &= U^n + \Delta t(1-2\gamma)f(U^*) + \Delta t\gamma f(U^{**}). \end{aligned}$$

Note that both equations are very similar to the equation to be solved at each step of the familiar implicit Euler scheme.

Only two values of the parameter γ will be considered:

(i) $\gamma = 1/2$. In this case U^{**} coincides with U^* and there is really only one system to solve per step. The method then reduces to the familiar midpoint rule

$$U^{n+1} = U^n + \Delta t f(\frac{1}{2}(U^n + U^{n+1}))$$

(see [18], Sect. 3.3.2). The method has order $p = 2$, is A-stable and conserves (see [5]) quadratic invariants such as the energy I_2.

(ii) $\gamma = (3+\sqrt{3})/6$, yielding a third order ($p = 3$), A-stable method discovered by Nørsett and Crouzeix (see [11], Chapter II, Table 7.2 and [12], Chapter III, Table 6.3). The method, which we denote by SDIRK3, does *not* conserve quadratic invariants.

The purpose of the experiments to be reported below is to illustrate the preceding theoretical results, rather than to establish a comparison between the practical performance of both integrators. For such a comparison to be meaningful, one should look at the errors yielded by the methods when both are using the same amount of computational time. However below we run the methods with equal values of Δt and this is biased in favour of the SDIRK3 integrator, which requires more work per step.

The validity of the hypotheses (H1)–(H2) may be established by standard analyses. On the other hand, it is easy to check that the leading terms of the local errors of the methods satisfy the requirements in (H3). Furthermore it is an exercise to show that for the midpoint rule both $< s, 1 >$ and $< s, w_0 >$ vanish, while for the SDIRK3 scheme $< s, 1 >= 0$ but $< s, w_0 >\neq 0$.

5.2. *One soliton experiments*

Our first group of experiments corresponds to the motion of a single soliton analyzed in Sect. 4. We fix the soliton parameters at the values $A_0 = 12$ (velocity $c_0 = 4$) and $\mu_0 = -10$ and integrate from $t = 0$ to $t = 10$. The soliton thus travels from $x = -10$ to $x = 30$. To implement in practice the semidiscrete schemes under consideration we discretize accurately the spatial variable on a fine grid so that all errors to be reported correspond to the time-stepping schemes. We compute the spatial derivatives by a Fourier pseudospectral approximation with 256 modes in the interval $-20 \le x \le 60$ (see, e.g., [6] for details). The errors introduced by the exponentially accurate space discretization are negligible; this was checked by refining the spatial grid.

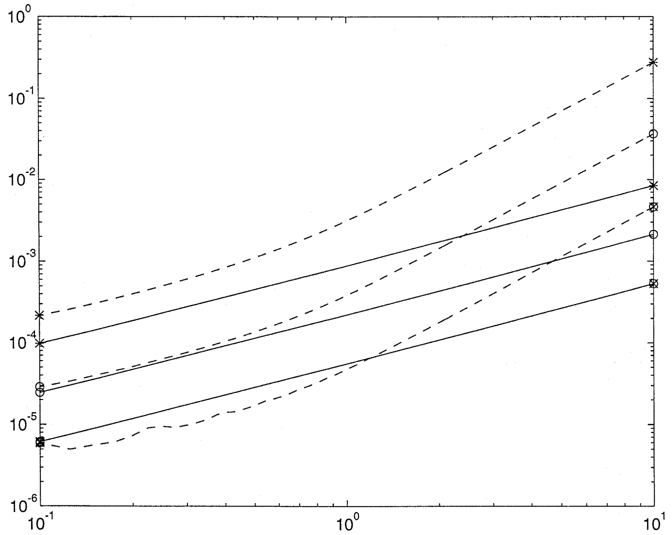

Fig. 1. L^2-error against t. Solid line: midpoint rule; broken line: SDIRK3. The time steps are $\Delta t = 1/40, 1/80, 1/160$

Figure 1 gives, in a log-log scale, the L^2-norm of the global error as a function of t. The solid lines correspond to the midpoint rule and the broken lines to SDIRK3. Shown are the runs corresponding to $\Delta t = 1/40$ ($\times$ signs), $1/80$ (circles) and $1/160$ ($\otimes$ signs). By examining the distance between the three parallel lines corresponding to a given method, we conclude that errors at a given value of t behave as Δt^2 for the midpoint rule and as Δt^3 for the SDIRK3 scheme. Thus these are accurate integrations where Δt is small enough (given the values of t) for the errors to show their expected asymptotic order. Further confirmation can be obtained from Table 1 that presents the errors at the final time $t = 10$. The slopes of the lines in Fig. 1 reveal that for the midpoint rule errors grow like t (cf. Theorem 2), while for SDIRK3 they grow like t^2 (cf. Theorem 1).

For any fixed value of t and small enough Δt the third order scheme will give smaller errors than the second order scheme. In Fig. 1 we see that, at $t = 1$,

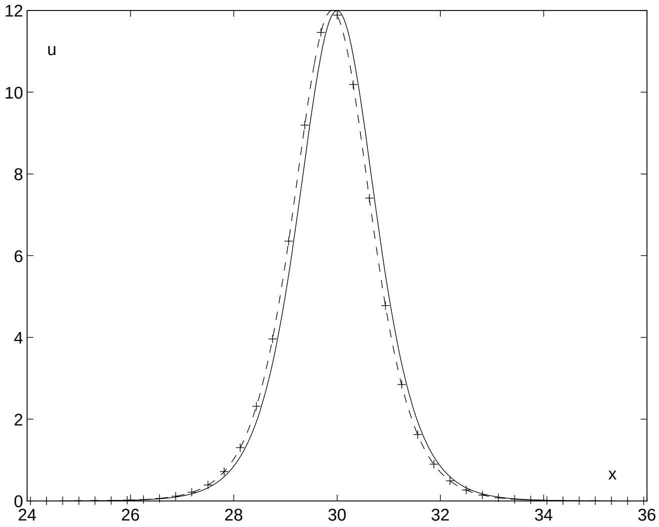

Fig. 2. Midpoint rule with $\Delta t = 1/40$ at $t = 10$. True soliton (solid line), modified soliton (broken line) and numerical result (crosses)

the crossover value of Δt is around $1/160$ and then the size of the error is approximately 5×10^{-5}. Since the norms of the errors behave as $K_1 t \Delta t^2$ and $K_2 t^2 \Delta t^3$, the crossover value of Δt behaves like $(K_1/K_2)t^{-1}$. Hence at $t = 10$ the crossover value would be of the order of 5×10^{-4}, when both methods would yield errors of about 5×10^{-6}. We conclude that, when integrating up to $t = 10$ or beyond with realistic values of Δt, it is not advisable to use the third order scheme.

Let us now investigate in detail the structure of the error, beginning with the midpoint rule. We have computed the coefficient λ_1 that features in the modified soliton in (33) and the L^2-norm of the difference between the computed solution U^n and the modified soliton. According to (33) we are then measuring the size of the complementary error plus the remainder. The results at the final $t = 10$ are displayed in Table 1. Two things should be noted. First, the errors with respect to the modified soliton are two orders of magnitude smaller than the true errors ('computed minus true'). In other words, the bulk of the error in the numerically computed soliton corresponds to the phase shift $\Delta t^2 \lambda_1 t$. Thus the numerically computed soliton possesses essentially a true KdV soliton profile (14) and keeps the right amplitude A_0 while travelling at an erroneous speed $c_0 - \Delta t^2 \lambda_1$. This is confirmed by Fig. 2 that shows, at $t = 10$, the true soliton solution (solid line), the numerical solution (crosses) and the modified soliton ($\Delta t = 1/40$). In Fig. 3 we have again displayed the computed solution at $t = 10$ but now with a vertical scale blown-up by three orders of maginitude. We see that the computed solution consists of a numerical soliton along with a small amplitude oscillation. No soliton tail appears

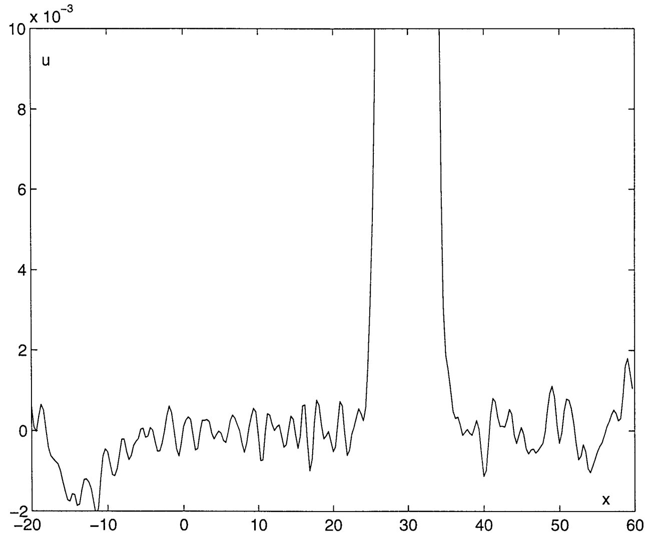

Fig. 3. Midpoint rule with $\Delta t = 1/40$ at $t = 10$. Numerical solution. The vertical scale has been magnified by three orders of magnitude

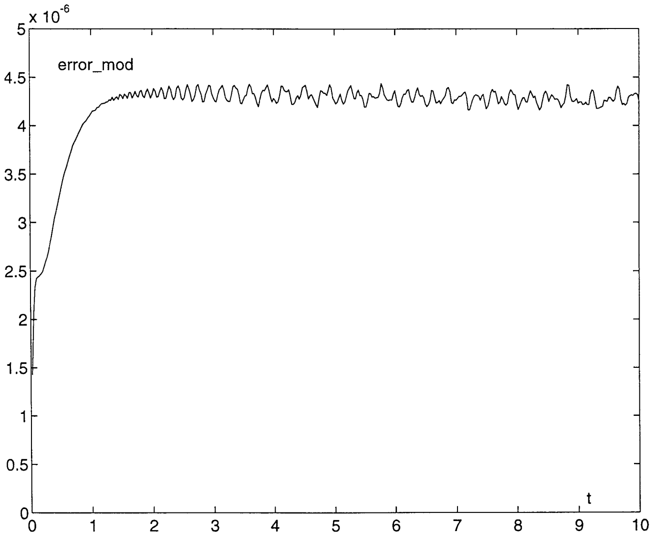

Fig. 4. L^2-error with respect to the modified soliton against t. Midpoint rule with $\Delta t = 1/160$

Table 1.

	Midpoint		SDIRK3	
Δt	Error	Error-mod	Error	Error-mod
2.50E-2	8.44E-3	6.45E-5	2.76E-1	1.69E-2
1.25E-2	2.12E-3	1.59E-5	3.66E-2	5.04E-4
6.25E-3	5.31E-4	4.26E-6	4.62E-3	3.62E-5

The second thing to be noted in Table 1 is that the errors with respect to the modified soliton show an $O(\Delta t^2)$ behaviour. This indicates that, for the small values of Δt considered, the complementary error dominates over the higher order remainder. In fact for the midpoint rule, the remainder $\Delta t\tilde{R}$ is expected to be specially small. It should behave as $o(\Delta t^3)$ rather than as $o(\Delta t^2)$: since the method is symmetric only even powers of Δt appear in the expansion of the global error ([11], Chapter II, Theorem 8.10). In Fig. 4 we have plotted the norm of the error with respect to the modified soliton for $\Delta t = 1/160$. As we have just discussed, this essentially corresponds to the norm of the complementary error $\Delta t^2\rho$. The bounded behaviour predicted by (34) is clearly borne out. The complementary error here is due to the numerically computed soliton not having exactly a true KdV profile (1). In the figure we see that at an initial transient regime the soliton shape evolves from the true KdV shape it possesses at $t = 0$ to the 'numerical' shape. In this transient, the complementary error builds up. After the transient, there is no further change in the soliton shape and the complementary error stops growing. This concludes our study of the midpoint rule results in the one soliton solution.

For the SDIRK3 method, Fig. 5 shows, at $t = 10$ the true soliton solution (solid line), the numerical solution (crosses) and the modified soliton ($\Delta t = 1/40$). We see that, within plotting accuracy, the computed solution coincides with the modified soliton, i.e., the complementary error and the remainder are, again, negligible. Now, according to Theorem 1, there are two sources of error implicit in the modified soliton: a small $O(t\Delta t^3)$ amplitude damping ($\lambda_2 < 0$) and a comparatively large $O(t^2\Delta t^3)$ phase shift. Both effects are clearly visible in Fig. 5.

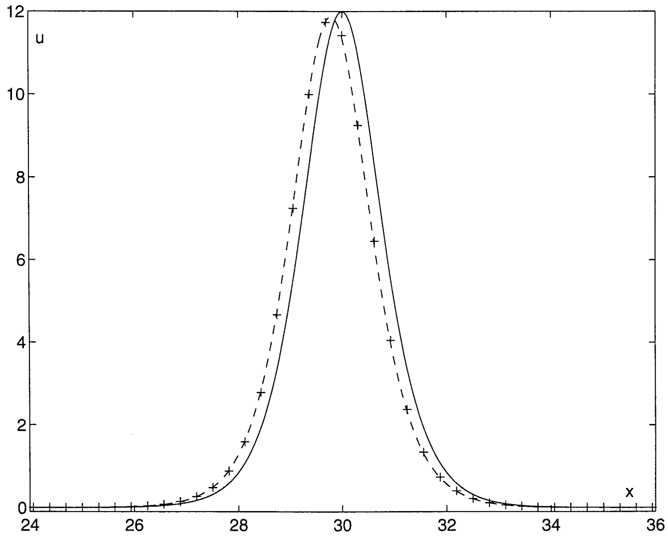

Fig. 5. SDIRK3 with $\Delta t = 1/40$ at $t = 10$. True soliton (solid line), modified soliton (broken line) and numerical result (crosses)

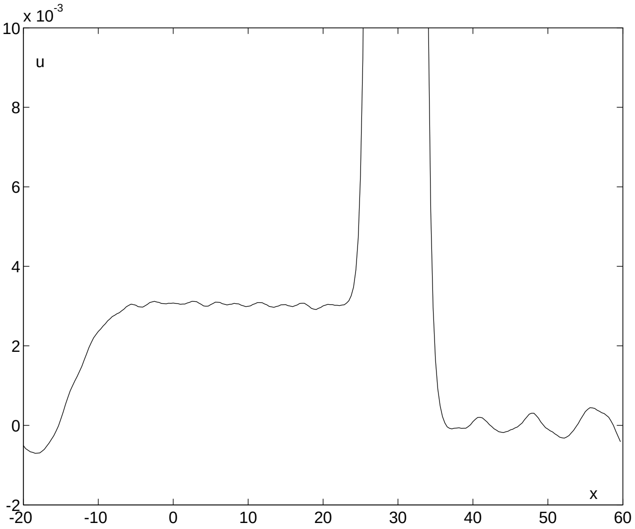

Fig. 6. SDIRK3 with $\Delta t = 1/40$ at $t = 10$. Numerical solution. The vertical scale has been magnified by three orders of magnitude

In Fig. 6 we have again displayed the computed solution at $t = 10$, but now with a vertical scale blown up by three orders of magnitude. A soliton 'tail' is apparent (cf. Fig. 3). The height of the tail is of about 3×10^{-3} and does not increase with t. The tail 'begins' at the location $x = \mu_0$ where the soliton started and 'ends' at the current soliton position. The existence of the tail and the value of the height can be predicted by soliton perturbation theory (see e.g. [14]), because, as shown by (7) and (19), numerical discretization amounts to a perturbation of the differential equation. The mechanism leading to the development of the tail is as follows. The amplitude of the modified soliton keeps decaying at a rate $|\Delta t^3 \lambda_2/3|$. Hence (cf. (16)) the modified soliton keeps losing its mass I_1. Since the numerical solution U^n is mass conserving ($< s, 1 >= 0$), the mass lost in the modified soliton must be gained somewhere. In fact the constant tail height is such that the mass in the trailing tail increases at the same rate as the mass in the modified soliton decreases (cf. [10]).

Figure 7 is a plot of error with respect to the modified soliton against time ($\Delta t = 1/160$). This includes the complementary and remainder errors. The linear growth predicted in (31) and due to the tail is apparent. The same plot at $\Delta t = 1/80$ or $1/40$ (not shown in the paper) shows, for t close to 10, an error growth higher than linear. This is explained as follows. An analysis similar to that leading to Theorem 1 reveals that the main part of the remainder $\Delta t \tilde{R}$ consists of an additional $O(t^2 \Delta t^4)$ phase shift. The complementary error, behaving as $O(t \Delta t^3)$, is likely to be hidden by the remainder when t is large and Δt moderate. In fact, the last column of Tab. 1 does not show an $O(\Delta t^3)$ behaviour, which suggests that, for the range of values of Δt considered, the remainder is not small relatively to the complementary error.

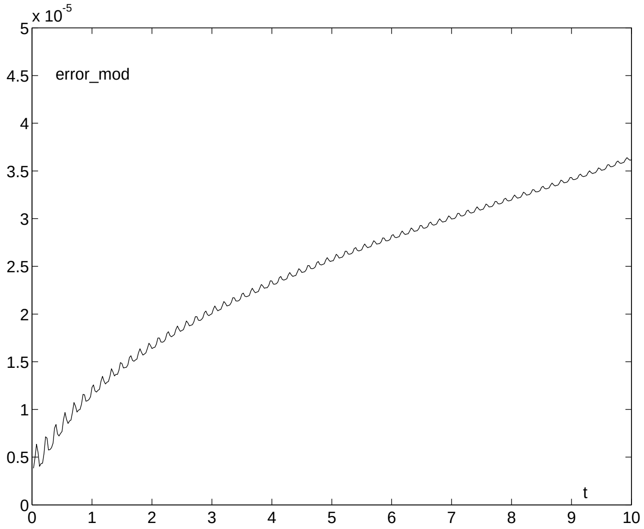

Fig. 7. L^2-error with respect to the modified soliton against t. SDIRK3 with $\Delta t = 1/160$

5.3. Soliton interaction

We have also performed experiments to investigate whether the advantages of conservation borne out in the preceding subsection and backed by the analysis in Sect. 4 for one soliton solutions also hold for more general solutions.

We studied the interaction between two solitons of amplitudes 12 and 6. We worked in a time interval $0 \leq t \leq 20$; this is long enough for the solitons to interact and to emerge from the interaction (for the single soliton solution we took a shorter interval $0 \leq t \leq 10$). We employed a spatial interval of length 130 and the pseudospectral method with 512 Fourier modes. This ensures that the spatial discretization errors are negligible. The grid spacing is now $130/512 \approx 0.25$, slightly smaller than the spacing $80/256 \approx 0.31$ used for the single soliton experiments. Figure 8 shows error versus time for the midpoint rule, $\Delta t = 1/40, 1/80, 1/160$. The error grows linearly before and after the interaction. For the SDIRK3 method (Fig. 9) the growth is quadratic. Note the different vertical scales in Figs. 8–9; the superiority of the conservative scheme is clear.

6. Conclusions and extensions

We have analyzed in detail the behaviour of the leading $O(\Delta t^p)$ term of the global error in the time integration of the KdV soliton. The most harmful component of that error is a quadratic phase error which is excited by the projection onto the energy gradient of the leading term of the local error. In the case of energy conserving schemes the quadratic growth is therefore absent. Conservation properties thus have a big impact on the global errors in the numerical

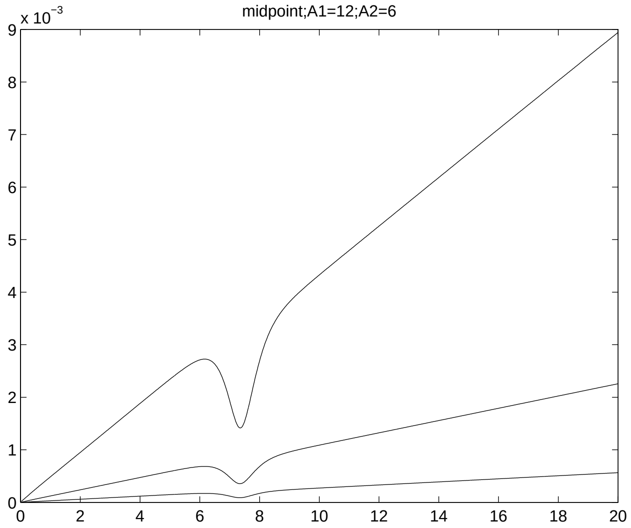

Fig. 8. Soliton interaction with the midpoint rule. L^2-error against t. The time steps are $\Delta t = 1/40, 1/80, 1/160$

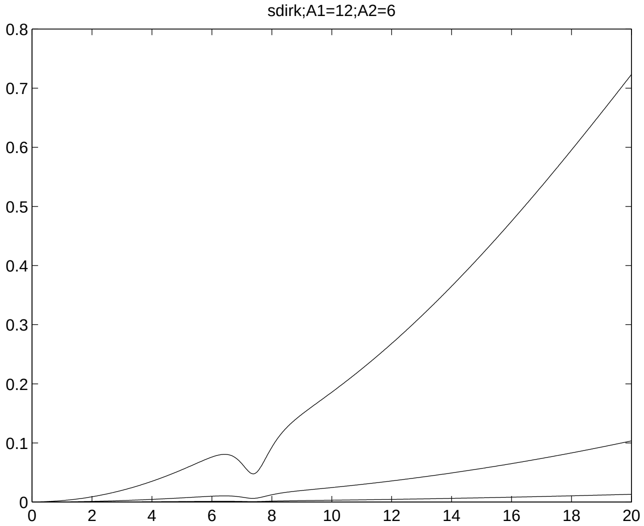

Fig. 9. Soliton interaction with SDIRK3. L^2-error against t. The time steps are $\Delta t = 1/40, 1/80, 1/160$

solution. An example was presented where a conserving second-order scheme gives, for all realistic values of Δt, more accurate results than a nonconserving, third-order scheme, in spite of the fact that the higher order method has smaller local (truncation) errors.

The error estimates presented here show linear or quadratic error growth. This should be compared with standard estimates that grow *exponentially* with t. This improvement is possible by restricting the attention to particular solutions (solitons) and carefully analyzing the corresponding variational equation.

Our analysis can be extended in many ways. It is straightforward to study the structure of the terms $O(\Delta t^{p+j})$, $j = 1, \dots, p-1$ of the expansion of the global error. All these terms satisfy the variational equation (19) with sources that, for energy-conseving methods, are orthogonal to the energy gradient, cf. Lemma 2.

As a second, relatively easy, generalization we could have considered fully discrete methods or methods with discrete x and continuous t: the error propagation equation for all those methods is still (19).

Finally, more general equations can be considered. The analysis by Pego and Weinstein [16], which is the basis of our results, applies to the case where the term uu_x is replaced by other nonlinearities of the form $f(u)_x$, so that catering for this more general case would have been a simple matter. We strongly believe that the extension to even more general families of equations having solitary wave solutions with amplitude-dependent velocity is possible. Of course the extension of our analysis would require a deep analysis of the linearized equations, as that carried out in [16] for the KdV-like case.

An interesting question raised by one of the referees is whether it is possible to construct schemes with $\langle s, 1\rangle = \langle s, w_0\rangle = \langle s, \eta_1\rangle = 0$. According to Theorem 2 those schemes would have, for a single soliton solution, a leading error term bounded in time.

Acknowledgement. This research has been supported by project DGICYT PB92–254.

References

1. Arakawa, A. (1966): Computational design for long-term numerical integration of the equations of fluid motion: two-dimensional incompressible flow. Part I. J. Comput. Phys. **1** 119–143
2. Benjamin, T.B. (1972): The stability of solitary waves. Proc. R. Soc. Lond. A. **328**, 153–183
3. Calvo, M.P., Sanz-Serna, J.M. (1993): The development of variable-step symplectic integrators, with application to the two-body problem. SIAM J. Sci. Comput. **14** 936–952
4. Cano, B., Sanz-Serna J.M.: Error growth in the numerical integration of periodic orbits, with application to Hamiltonian and reversible system. SIAM J. Numer. Anal. to appear
5. Cooper, G.J. (1987): Stability of Runge-Kutta methods for trajectory problems. IMA J. Numer. Anal. **7**, 1–13
6. de Frutos, J., Sanz-Serna, J.M. (1992): An easily implementable fourth-order method for the time integration of wave problems. J. Comput. Phys. **103**, 160–168
7. de Frutos J., Sanz-Serna, J.M. (1994): Erring and being conservative. In: Numerical Analysis 1993, Griffiths , D.F., Watson, G.A. (eds.), Longman Scientific and Technical, pp. 75–88
8. de Frutos, J., Sanz-Serna, J.M. (1993): Error growth and invariant quantities in numerical methods: a case study. Applied Mathematics and Computation Report 1993/4, Universidad de Valladolid, Spain

9. Gear, C.W. (1990): Invariants and numerical methods for ODEs. Physica D **60**, 303–310
10. Grimshaw, R., Mitsudera, H. (1993): Slowly varying solitary wave solutions of the perturbed Korteweg-de Vries equation revisited. Studies Appl. Maths. **90**, 75–86
11. Hairer, E., Nørsett, S.P., Wanner, G. (1993): Solving Ordinary Differential Equations I, Nonstiff problems, 2nd edn. Springer, Berlin, 1993
12. Hairer, E., Wanner, G. (1991): Solving Ordinary Differential Equations II, Stiff and Differential-Algebraic problems. Springer, Berlin, 1991
13. Karpman, V.I., Maslov, E.M. (1977): Perturbation theory for solitons. Sov. Phys. JETP **46**, 81–291
14. Karpman, V.I., Maslov, E.M. (1978): Structure of tails produced under the action of perturbations of solitons. Sov. Phys. JETP **48**, 252–259
15. Lax, P., Wendroff, B. (1960): Systems of conservation laws. Comm. Pure Appl. Math. **13**, 217–237
16. Pego, R.L., Weinstein, M.I. (1994): Asymptotic stability of solitary waves. Comm. Math. Phys. **164**, 305–349
17. Richtmyer, R.D., Morton, K.W. (1967): Difference Methods for Initial Value Problems. Wiley-Interscience, New York, 1967
18. Sanz-Serna, J.M., Calvo, M.P. (1994): Numerical Hamiltonian Problems. Chapman & Hall, London, 1994
19. Shampine, L.F. (1986): Conservation laws and the numerical solution of ODEs. Comp. & Maths with Appls. **12B**, 1287–1296

This article was processed by the author using the LaTeX style file *pljour1m* from Springer-Verlag.